ELSEVIER Linear Algebra and its Applications 275–276 (1998) 417–431

LINEAR ALGEBRA AND ITS APPLICATIONS

Spectral sensitivity of products of bidiagonals [1]

Beresford N. Parlett [2]

Mathematics Department and Computer Science Division, EECS Department, University of California, Berkeley, CA 94720, USA

Received 28 November 1996; accepted 18 August 1997

Submitted by V. Mehrmann

Abstract

Simple formulae are derived for certain relative condition numbers. Tridiagonal matrices may be represented as products of bidiagonals in various ways depending on properties such as symmetry and positive definiteness. The condition numbers give the amplification factor for relative changes in a nonzero eigenvalue caused by relative changes in an entry of a bidiagonal factor. The formulae show that in many, but not all cases these condition numbers are of modest size. Several examples illustrate the results and raise new questions.

1. Introduction

Any engineer who is aware of the limitations on his data must be interested in how well the output of some process is determined by the data. In this study the output is a set of eigenvalues and the input is usually a square matrix. Perturbation theory teaches us some hard lessons. In general some eigenvalues may be exceedingly sensitive to small changes in the matrix entries while others may be the opposite. For Hermitian matrices, and these include the real symmetric matrices, no eigenvalue can change by more than the norm of the change in the matrix (H. Weyl). See Chapter 10 of [1] for a simple proof. However, in some applications, the spectrum ranges over many orders of magnitude and it is usually the eigenvalues near 0 that are of interest. Unfortunately

[1] This research paper is based on a talk given at the minisymposium on Perturbation Theory during the 6th ILAS Conference in Chemnitz, Germany, in August 1996.

[2] Supported by ONR, Contract N000014-90-J-1372.

0024-3795/98/$19.00
PII: S0024-3795(97)10052-0

Weyl's theorem does not guarantee that these small eigenvalues have as many correct decimal digits as the large (unwanted) ones.

In a number of important cases a generalized Hermitian problem $(K - \lambda M)\boldsymbol{u} = \boldsymbol{0}$ does determine its small eigenvalues to high relative accuracy but this valuable property has been lost by heeding the mathematician's advice to reduce a difficult problem (say the pair K, M) to an easier one (a standard eigenproblem $A - \lambda I$ with $A = KM^{-1}$ or $A = M^{-1/2}KM^{-1/2}$). There is more on this topic in Section 4.

Even symmetric tridiagonal matrices that are positive definite do not always determine their tiny eigenvalues to an acceptable number of figures.

Armed with this sobering knowledge numerical analysts are careful not to try to force their algorithms to achieve unattainable accuracies.

The objective that is repeated, like a mantra, throughout the community of eigenvalue hunters is to achieve 'as much accuracy as the data warrants'.

We acknowledge clever recent research that overlaps the results given here. See [2–5], for example. The formulae in Theorem 1 are new but the 'result' is known, see [6]. Theorem 2 is not as general as results given in [5], but Theorem 3 does break new ground. We show that a little calculus gives a lot of insight. Kahan's ingenious proof, in 1966 (!) but see [6], that all singular values are determined to high relative accuracy by the entries in a bidiagonal matrix was not really necessary! We are not suggesting that traditional perturbation theory be replaced by calculus but people do remember simple approaches and we show how naturally a representation as a product of matrices can lead to expressions for relative accuracy.

Before starting let us note that a single matrix M may be regarded as a representative of the equivalence class of all its translates $\{M + \sigma I, \sigma \in \mathbb{C}\}$. In this sense there is no natural origin for the standard eigenvalue problem. However it is not so easy to represent in factored form the translates of a product FG of invertible matrices.

2. Background

The adjectives absolute and relative will occur frequently in the investigation and it seems advisable to say what is meant.

Let B be a square matrix that depends smoothly on some parameter τ. Following Newton's notation for derivatives let $\dot{B}$ denote the derivative of B with respect to τ and consider a typical simple eigenvalue λ with column and row eigenvectors $\boldsymbol{x}$ and $\boldsymbol{y}^*$. Thus

$$B\boldsymbol{x} = \boldsymbol{x}\lambda, \qquad \boldsymbol{y}^*B = \lambda\boldsymbol{y}^*,$$

and without loss of generality one may assume $\boldsymbol{y}^*\boldsymbol{x} > 0$. Standard differentiation yields

$$\dot{\lambda} = \frac{\boldsymbol{y}^* \dot{B} \boldsymbol{x}}{\boldsymbol{y}^* \boldsymbol{x}} = \frac{\|\boldsymbol{y}^*\| \cdot \|\boldsymbol{x}\|}{\boldsymbol{y}^* \boldsymbol{x}} \cdot \frac{\boldsymbol{y}^* \dot{B} \boldsymbol{x}}{\|\boldsymbol{y}^*\| \cdot \|\boldsymbol{x}\|}$$

and $\|\boldsymbol{y}^*\| \cdot \|\boldsymbol{x}\| / \boldsymbol{y}^* \boldsymbol{x}$ is the (absolute) condition number of λ; cond(λ). In particular the sensitivity of λ to a change in b_{jk}, i.e $\tau = b_{jk}$, is given by

$$\frac{\partial \lambda}{\partial b_{jk}} = \frac{\boldsymbol{y}^* \boldsymbol{e}_j \cdot \boldsymbol{e}_k^* \boldsymbol{x}}{\boldsymbol{y}^* \boldsymbol{x}} = \frac{\bar{y}(j) x(k)}{\boldsymbol{y}^* \boldsymbol{x}},$$

where $I = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$.

We call $|\partial \lambda / \partial b_{jk}|$ the *absolute* sensitivity (of λ to b_{jk}). We may also be interested in the *relative* sensitivity of λ, for $\lambda \neq 0$, i. e. $|(\partial \lambda / \partial b_{jk})/\lambda|$. Even more relevant is the relative change in λ due to small relative changes in the (j,k) entry, i.e.

$$\left| \frac{\partial \lambda}{\partial b_{jk}} \cdot \frac{b_{jk}}{\lambda} \right|$$

and this is what we mean by the *relative sensitivity*. If $\left|(\partial \lambda / \partial b_{jk}) \cdot (b_{jk}/\lambda)\right| = 10^2$ we would say, somewhat loosely, that uncertainty in the 5th decimal digit of b_{jk} provokes uncertainty in the 3rd decimal digit of λ. The quantity cond(λ) defined above is absolute and it is an upper bound on the sensitivity of λ to (absolute) changes in any entry. As mentioned in the introduction, there are examples of positive definite symmetric tridiagonal matrices such that *relative* changes to some off diagonal entries of order ϵ produce *relative* changes in some eigenvalues of order $\sqrt{\epsilon}$. Thus we cannot be sure, in general, that tiny λ are determined to any correct figures when matrix entries are uncertain or noisy. This is sad but true and prompts us to look for alternative representations of matrices that might determine all eigenvalues well, regardless of magnitude.

Recent investigations have shown that certain classes of matrices do determine their small eigenvalues to high relative accuracy. See [6–10].

3. Singular values of bidiagonals

In 1966, W. Kahan proved a surprising result which is in stark contrast to the high relative sensitivities of small eigenvalues to small changes in matrix entries. He showed that small relative changes in the entries of a bidiagonal matrix lead to small relative changes in all the singular values, see [6]. Let

$$B = \operatorname{bidiag}\begin{pmatrix} b_1 & b_2 & \cdot & b_{n-2} & b_{n-1} \\ a_1 & a_2 & \cdot & \cdot & a_{n-1} & a_n \end{pmatrix}$$

have $(\sigma, \boldsymbol{v}, \boldsymbol{u})$ as a typical singular triple:

$$Bv = u\sigma, \qquad B^t u = v\sigma, \tag{1}$$

$$v^t v = u^t u = 1. \tag{2}$$

Suppose that $a_i > 0$, $i = 1, 2, \ldots, n$, and $b_j > 0$, $j = 1, \ldots, n-1$.

To show that eigenvalues have not been abandoned note that if

$$\frac{\partial \sigma}{\partial a_k} \cdot \frac{a_k}{\sigma} = \kappa,$$

then this is equivalent to a result concerning $\lambda = \sigma^2$ of $B^t B$ since

$$\frac{\partial \lambda}{\partial a_k} \cdot \frac{a_k}{\lambda} = 2\sigma \frac{\partial \sigma}{\partial a_k} \cdot \frac{a_k}{\sigma^2} = 2\kappa.$$

Kahan's original proof was not published and uses Sylvester's Inertia theorem in an ingenious way. The proof in [6] is a little more general but follows the same far from obvious reasoning. A simple proof is given in [11].

In contrast our Theorem 1 is straightforward and yields stronger results than Theorem 2 in [6] for tiny perturbations. In our terminology, Kahan bounded the relative condition number by 1 but that bound holds for all relative perturbations both large and small while our Theorem 1 applies only to infinitessimal changes.

In what follows it is convenient to define

$$b_0 = b_n = 0.$$

Theorem 1. *With the notation given in Eqs. (1) and (2), since* $\sigma \neq 0$,

$$(a) \qquad \frac{\partial \sigma}{\partial a_k} \cdot \frac{a_k}{\sigma} = \sum_{i=1}^{k} v(i)^2 - \sum_{j=1}^{k-1} u(j)^2 = \sum_{m=k}^{n} u(m)^2 - \sum_{l=k+1}^{n} v(l)^2,$$

$$(b) \qquad \frac{\partial \sigma}{\partial b_k} \cdot \frac{b_k}{\sigma} = \sum_{i=1}^{k} \left(v(i)^2 - u(i)^2 \right) = \sum_{m=k+1}^{n} \left(u(m)^2 - v(m)^2 \right).$$

Proof. First derive well known expressions for $\partial\sigma/\partial a_k$ and $\partial\sigma/\partial b_k$. From Eqs. (1) and (2),

$$\sigma = u^t B v. \tag{3}$$

Write out Eq. (3) using B's entries to find

$$\frac{\partial \sigma}{\partial a_j} = u(j)v(j), \qquad \frac{\partial \sigma}{\partial b_j} = u(j)v(j+1). \tag{4}$$

These expressions are simple but do not reveal the dependence on σ, a_j, and b_j. To remedy the situation write out each equation in (1) in detail

$$a_j v(j) + b_j v(j+1) = u(j)\sigma, \qquad b_{j-1} u(j-1) + a_j u(j) = v(j)\sigma. \tag{5}$$

Multiply Eq. (4) by a_j and b_j and substitute into Eq. (5) to find, for $j = 1, 2, \ldots, n$,

$$a_j \frac{\partial \sigma}{\partial a_j} + b_j \frac{\partial \sigma}{\partial b_j} = u(j)^2 \sigma, \qquad b_{j-1} \frac{\partial \sigma}{\partial b_{j-1}} + a_j \frac{\partial \sigma}{\partial a_j} = v(j)^2 \sigma. \tag{6}$$

Now set $j = 1, 2, \ldots, n$, in turn, to find

$$\begin{aligned}
a_1 \frac{\partial \sigma}{\partial a_1} &= \sigma v(1)^2, \\
b_1 \frac{\partial \sigma}{\partial b_1} &= \sigma\left(u(1)^2 - v(1)^2\right), \\
a_k \frac{\partial \sigma}{\partial a_k} &= \sigma\left[\sum_{i=1}^{k} v(i)^2 - \sum_{j=1}^{k-1} u(i)^2\right], \\
b_k \frac{\partial \sigma}{\partial b_k} &= \sigma \sum_{i=1}^{k}\left(u(i)^2 - v(i)^2\right).
\end{aligned}$$

Since $\boldsymbol{u}^t\boldsymbol{u} = \boldsymbol{v}^t\boldsymbol{v} = 1$, the complementary expressions follow from $\sum_{i=1}^{k} v(i)^2 = 1 - \sum_{j=k+1}^{n} v(j)^2$. In particular, $a_n \partial\sigma/\partial a_n = \sigma u(n)^2$. □

Corollary 1. *For* $k = 1, 2, \ldots, n$,

$$\left|\frac{\partial \sigma}{\partial a_k} \cdot \frac{a_k}{\sigma}\right| < 1, \qquad \left|\frac{\partial \sigma}{\partial b_k} \cdot \frac{b_k}{\sigma}\right| < 1,$$

$$0 \leqslant \frac{\partial \sigma}{\partial a_k} \cdot \frac{a_k}{\sigma} + \frac{\partial \sigma}{\partial b_k} \cdot \frac{b_k}{\sigma} = u(k)^2 < 1.$$

Proof. Expressions of the form $\sum v(i)^2 - \sum u(j)^2$ must lie in $[-1, +1]$. The assumption that $a_k b_k \neq 0$, $k = 1, \ldots, n-1$, ensures that equality is never obtained, e. g. $|v(1)| < 1$. □

The formulae in Theorem 1 show that we may have

$$\frac{\partial \sigma}{\partial a_k} \cdot \frac{a_k}{\sigma} \ll 1, \qquad \frac{\partial \sigma}{\partial b_k} \cdot \frac{b_k}{\sigma} \ll 1.$$

Examples of very small relative condition numbers are given in Section 6 and show that this phenomenon is common and is not confined to the positive definite case.

Corollary 2. *The eigenvalues of a symmetric tridiagonal matrix that is positive definite are determined to high relative accuracy by the entries in the Cholesky factor.*

4. A classical example

A system of masses connected by springs is a standard example in courses on dynamics. The equations of motion $M\ddot{x} + Kx = \mathbf{0}$ give rise to an eigenvalue problem $(K - \omega^2 M)v = \mathbf{0}$ with $M = \mathrm{diag}(m_1, \ldots, m_n)$, $x = v\exp(\mathrm{i}\omega t)$ and

$$K = \mathrm{tridiag}\begin{pmatrix} & -k_2 & & -k_3 & \cdot & -k_n & \\ k_1 + k_2 & & k_2 + k_3 & & \cdot & & k_n \\ & -k_2 & & -k_3 & \cdot & -k_n & \end{pmatrix}$$

where the k_i are the spring constants. The point is that K may be factored as

$$K = BK_{\mathrm{d}}B^{\mathrm{t}},$$

where $K_{\mathrm{d}} = \mathrm{diag}(k_1, \ldots, k_n)$ and

$$B = \mathrm{bidiag}\begin{pmatrix} -1 & -1 & \cdot & -1 & -1 & \\ 1 & 1 & \cdot & \cdot & 1 & 1 \end{pmatrix}.$$

It follows that the desired frequencies ω satisfy

$$\left[\left(M^{-1/2}BK_{\mathrm{d}}^{1/2}\right)\left(M^{-1/2}BK_{\mathrm{d}}^{1/2}\right)^{\mathrm{t}} - \omega^2 I\right]\boldsymbol{u} = \mathbf{0}.$$

So the ω are the singular values of $U = M^{-1/2}BK_{\mathrm{d}}^{1/2}$ and each ω, however small, is determined to high relative accuracy by the entries of U. In contrast the small ω^2 are not always determined to high relative accuracy by the entries of $M^{-1/2}KM^{-1/2}$ where the additions on the diagonal discard important information. The tactical error, passed down from teacher to student for decades, was to use K instead of $BK_{\mathrm{d}}B^{\mathrm{t}}$, the product form. There are no adds or subtracts in forming U whose jth row contains $a_j = (k_j/m_j)^{1/2}$, $b_j = -(k_{j+1}/m_j)^{1/2}$. See [12]. Current research considers whether the B, K_{d}, and M matrices that arise in 2D and 3D applications also define the small eigenvalues to high relative accuracy.

5. Symmetric tridiagonals

If T is not positive definite (p.d.) but $T + \sigma I$ is p.d. then Theorem 1 permits us to say that the quantities $(\lambda + \sigma)$ are determined to high relative accuracy by the Cholesky factor of $T + \sigma I$. This is not very satisfactory and so the next step is to see to what extent Theorem 1 can be extended.

All tridiagonals T may be written as

$$T = L\Omega L^{\mathrm{t}}, \tag{7}$$

where L is lower triangular with positive diagonal and Ω is a signed, symmetric permutation matrix. Ω will be a direct sum of ± 1 and

$$\pm \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

entries. Note that $\Omega^2 = I$. The spectral factorization of invertible T is written as

$$T = S\Lambda S^{\mathrm{t}}, \qquad I = SIS^{\mathrm{t}} \tag{8}$$

and each eigenvalue λ is written as

$$\lambda = \operatorname{sign}(\lambda)\sigma^2, \tag{9}$$

so that $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ is positive definite. Note that Eq. (8) is unique only up to permutations of Λ, i.e. the labelling of eigenvalues.

Our interest is in how the eigenvalues of $T = L\Omega L^{\mathrm{t}}$ respond to small relative changes to the nonzero entries of L while Ω remains fixed. Theorem 1 used the singular vectors of L to express the sensitivity when $\Omega = I$. When $\Omega \neq I$ it is useful to introduce an indefinite, or improper, SVD of L, namely

$$L = S\Sigma P^{\mathrm{t}}, \tag{10}$$

where S and Σ are given by Eqs. (8) and (9) and P is Ω-sign(Λ)-orthogonal, i.e.

$$P^{\mathrm{t}}\Omega P = \operatorname{sign}(\Lambda). \tag{11}$$

When some $\Omega \neq I$ then Ω defines an indefinite 'inner product' $\langle \boldsymbol{u}, \boldsymbol{v} \rangle := \boldsymbol{v}^{\mathrm{t}}\Omega\boldsymbol{u}$ that satisfies all the axioms except for positivity.

From Eq. (10),

$$P = L^{\mathrm{t}}S\Sigma^{-1}, \tag{12}$$

we find

$$\begin{aligned} P^{\mathrm{t}}\Omega P &= \Sigma^{-1}S^{\mathrm{t}}L\Omega L^{\mathrm{t}}S\Sigma^{-1} \\ &= \Sigma^{-1}S^{\mathrm{t}}TS\Sigma^{-1} \\ &= \operatorname{sign}(\Lambda), \quad \text{by Eq. (9)}, \end{aligned}$$

verifying Eq. (11).

Theorem 2. *An invertible symmetric tridiagonal matrix T may be written as $T = L\Omega L^{\mathrm{t}}$ with*

$$L = \operatorname{bidiag}\begin{pmatrix} a_1 & a_2 & \cdot & a_{n-1} & a_n \\ & b_1 & \cdot & b_{n-1} & \end{pmatrix},$$

$\Omega =$ *a signed symmetric perturbation matrix.*

Let the spectral factorization of T be given by Eqs. (8) and (9). Let the columns of $P = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n)$, defined by Eq. (12), be the right Ω-singular vectors of L and note that the columns of S are the left singular vectors of L.

Let $(\sigma, \mathbf{s}^t, \Omega\mathbf{p})$ *be a typical improper singular triple of* L *and let the corresponding eigenvalue of* T *be* $\operatorname{sign}(\lambda)\sigma^2$. *Let* $\Omega p(k)$ *denote the* k*th entry of* $\Omega\mathbf{p}$, *then*

$$\frac{a_1}{\sigma}\cdot\frac{\partial\sigma}{\partial a_1} = s(1)^2,$$

$$\frac{b_1}{\sigma}\cdot\frac{\partial\sigma}{\partial b_1} = \operatorname{sign}(\lambda)p(1)\Omega p(1) - s(1)^2,$$

$$\frac{a_k}{\sigma}\cdot\frac{\partial\sigma}{\partial a_k} = \sum_{i=1}^{k} s(i)^2 - \operatorname{sign}(\lambda)\sum_{j=1}^{k-1} p(j)\Omega p(j)$$

$$= \operatorname{sign}(\lambda)\sum_{i=k}^{n} p(i)\Omega p(i) - \sum_{j=k+1}^{n} s(j)^2,$$

$$\frac{b_k}{\sigma}\cdot\frac{\partial\sigma}{\partial b_k} = \operatorname{sign}(\lambda)\sum_{i=1}^{k} p(i)\Omega p(i) - \sum_{j=1}^{k} s(j)^2$$

$$= \sum_{i=k+1}^{n} s(i)^2 - \operatorname{sign}(\lambda)\sum_{j=k+1}^{n} p(j)\Omega p(j),$$

$$\frac{a_n}{\sigma}\cdot\frac{\partial\sigma}{\partial a_n} = \operatorname{sign}(\lambda)p(n)\Omega p(n).$$

Proof. Multiply each side of $L = S\Sigma P^t$ by ΩP and use Eq. (11) to find $L\Omega P = S\Sigma\ \operatorname{sign}(\Lambda)$, or

$$L\Omega\mathbf{p}_i = \mathbf{s}_i\sigma_i\operatorname{sign}(\lambda_i), \quad i = 1, 2, \ldots, n. \tag{13}$$

From the definition of L

$$L^t\mathbf{s}_j = P\Sigma S^t\mathbf{s}_j = P\Sigma\mathbf{e}_j = \mathbf{p}_j\sigma_j, \quad j = 1, 2, \ldots, n. \tag{14}$$

By Eq. (11),

$$\mathbf{p}_j^t\Omega\mathbf{p}_j = \operatorname{sign}(\lambda_j), \qquad \mathbf{s}_i^t\mathbf{s}_i = 1. \tag{15}$$

Premultiply Eq. (13) by $\mathbf{s}_i^t$, write out the expression $\operatorname{sign}(\lambda_i)\sigma_i = \mathbf{s}_i^tL\Omega\mathbf{p}_i$ to see that it is linear in a_j and b_j. Drop the index i to find

$$\operatorname{sign}(\lambda)\frac{\partial\sigma}{\partial a_j} = \Omega p(j)s(j), \qquad \operatorname{sign}(\lambda)\frac{\partial\sigma}{\partial b_j} = \Omega p(j)s(j+1). \tag{16}$$

Now use the bidiagonal form of Eqs. (13) and (14). For $b_0 = b_n = 0$ and $j = 1, 2, \ldots, n$,

$$b_{j-1}\Omega p(j-1) + a_j\Omega p(j) = \operatorname{sign}(\lambda)s(j)\sigma,$$
$$a_j s(j) + b_j s(j+1) = p(j)\sigma.$$

Multiply by $s(j)$ and $\Omega p(j)$, respectively, and use Eq. (16) to obtain

$$\operatorname{sign}(\lambda)b_{j-1}\frac{\partial\sigma}{\partial b_{j-1}} + \operatorname{sign}(\lambda)a_j\frac{\partial\sigma}{\partial a_j} = \operatorname{sign}(\lambda)s(j)^2\sigma,$$

$$\operatorname{sign}(\lambda)a_j\frac{\partial\sigma}{\partial a_j} + \operatorname{sign}(\lambda)b_j\frac{\partial\sigma}{\partial b_j} = p(j)\Omega p(j)\sigma.$$

Now set $j = 1, 2, \ldots, n$ in turn to obtain the conclusion of Theorem 2. □

Remark 1. From Eq. (9), it follows that, for $\lambda \neq 0$,

$$\frac{a_j}{\lambda}\frac{\partial\lambda}{\partial a_j} = 2\frac{a_j}{\sigma}\frac{\partial\sigma}{\partial a_j}, \qquad \frac{b_j}{\lambda}\frac{\partial\lambda}{\partial b_j} = 2\frac{b_j}{\sigma}\frac{\partial\sigma}{\partial b_j}.$$

Remark 2. From Eq. (12), $\boldsymbol{p} = L^{\mathrm{t}}\boldsymbol{s}/\sigma$ and $\|\boldsymbol{p}\| \leqslant \|L\|/\sigma$ but this inequality is rather crude. Since $L\Omega L^{\mathrm{t}}\boldsymbol{s} = \boldsymbol{s}\lambda$ an alternative expression for $\boldsymbol{p}$ is

$$\boldsymbol{p} = \Omega L^{-1}\boldsymbol{s} \cdot \sigma \cdot \operatorname{sign}(\lambda)$$

and so

$$\|\boldsymbol{p}\| \leqslant \sigma\|L^{-1}\|.$$

Thus

$$\|\boldsymbol{p}\| \leqslant \sqrt{\operatorname{cond}(L)} = \sqrt{\|L\| \cdot \|L^{-1}\|}.$$

But even this bound ignores cancellation in $L^{\mathrm{t}}\boldsymbol{s}$ and $L^{-1}\boldsymbol{s}$.

Remark 3. Section 4 indicates that there are important applications where the matrix of interest is given in product form and the small eigenvalues may be defined to high relative accuracy by the factors but not by their product. When a given product does not yield the desired accuracy it is recommended to look for better product representations and that is focus of our current investigations.

The message of Theorem 2 is that L certainly determines an eigenvalue λ of $L\Omega L^{\mathrm{t}}$ to high relative accuracy when the associated $\boldsymbol{p}$-vector has entries bounded by a modest value such as 10.

The point to notice is that when $\Omega = I$ the vector $\boldsymbol{p} = \Omega L^{-1}\boldsymbol{s}\sigma$ satisfies $\|\boldsymbol{p}\| = 1$ however ill-conditioned L may be. Thus we may expect that, even when $\Omega \neq I$, the bound $\|\boldsymbol{p}\| \leqslant \sqrt{\operatorname{cond}(L)}$ is far from tight in many, but not all, cases. The next section supplies some evidence.

6. Examples

Description of the tables given below. For each $\lambda \neq 0$ of each $n \times n$ tridiagonal T there are $2n - 1$ relative condition numbers (rcond), $\left|(a_j/\lambda)(\partial\lambda/\partial a_j)\right|$,

$j = 1, \ldots, n$, and $|(b_j/\lambda)(\partial\lambda/\partial b_j)|$, $j = 1, \ldots, n-1$. Here the a's and b's are the nontrivial entries of L. All this is too much information so we have selected the largest and smallest of these $2n-1$ magnitudes. To indicate the index (in the column called 'w.r.t.' meaning 'with respect to') $-j$ denotes b_j and j denotes a_j. Thus in Case (1) the largest condition number was for b_{39} and the smallest was for a_{40}.

Four eigenvalues were chosen for each matrix. Rows 1 and 2 show the eigenvalues nearest 0. Row 3 exhibits the most sensitive eigenvalue and Row 4 the least sensitive. Hence the repetitions.

(1) Consider L with 1's on the diagonals, 2's on the off-diagonals.

$$n = 40, \qquad \|L\| = 2.998, \qquad \mathrm{cond}(L) = 2.198 \times 10^{12},$$

$$\Omega = \mathrm{diag}(\,+\,+\,-\,-\,-\,+\,+\,+\,+\,+\,+\,+\,+\,+\,+\,+\,+\,+\,+\,+\,+\,+\,-\,+\,-\,+\,+\,-\,++$$
$$+\,+\,+\,+\,+\,+\,+\,+\,+\,+\,+\,+\,++).$$

λ	$\|\boldsymbol{p}\|$	max rcond	w.r.t	min rcond	w.r.t.
-1.39×10^{-16}	11.6	2	−39	2×10^{-8}	40
1.07	1	0.472	−13	7.06×10^{-7}	2
-1.39×10^{-16}	11.6	2	−39	2×10^{-8}	40
7.06	1	0.16	−39	0.1952×10^{-10}	3

Here L is chosen to be ill-conditioned without having any small entries. Its determinant is 1. The tiny eigenvalue is the most sensitive and yet it has relative condition numbers less than 2.

(2a) T is the [1,2,1] matrix shifted by $\mathrm{eig}(20) = 1.85054$

$$n = 41, \qquad \|L\| = 2.983 \times 10^{7}, \qquad \mathrm{cond}(L) = 3.282 \times 10^{15},$$

$$\Omega = \mathrm{diag}(\,+\,-\,+\,-\,+\,-\,+\,-\,+\,-\,+\,-\,+\,-\,+\,-\,+\,-\,+\,+\,-\,+\,-\,+\,-\,+\,-$$
$$+\,-\,+\,-\,+\,-\,+\,-\,+\,-\,+\,-\,+\,+).$$

λ	$\|\boldsymbol{p}\|$	max rcond	w.r.t	min rcond	w.r.t.
8.29×10^{-16}	1.06	1.906	40	0.0948	1
0.148	1.69×10^{7}	2.86×10^{14}	−20	2.66×10^{-15}	41
0.148	1.69×10^{7}	2.86×10^{14}	−20	2.66×10^{-15}	41
2.13	1.09	0.0958	13	4.12×10^{-18}	41

The matrix is as close to singular as we can make it. The smallest eigenvalue is detemined to very high relative accuracy but the second eigenvalue is extremely sensitive.

(2b) T is the [1,2,1] matrix shifted by (eig(20)+eig(21))/2 = 1.9252.

$$n = 41, \qquad \|L\| = 7.207, \qquad \mathrm{cond}(L) = 335.5,$$

$$\Omega = \mathrm{diag}(+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+)$$

λ	$\|\boldsymbol{p}\|$	max rcond	w.r.t	min rcond	w.r.t.
0.0747	8.74	33.2	41	3.22×10^{-15}	4
−0.0747	8.74	33.0	−40	2.72×10^{-14}	21
0.0747	8.74	33.2	41	3.22×10^{-15}	4
2.07	1.11	0.0994	−17	0.000474	40

In contrast to (2a) all eigenvalues are relatively robust and this is guaranteed by the small value of cond(L). Recall that $\|\boldsymbol{p}\| \leqslant \sqrt{\mathrm{cond}}(L)$.

(2c) T is the [1,2,1] matrix shifted by eig(5) = 0.13825

$$n = 41, \qquad \|L\| = 3.164, \qquad \mathrm{cond}(L) = 2.336 \times 10^{8},$$

$$\Omega = \mathrm{diag}(+++++++-++++++++-+++++++++-++++++-++++++++).$$

λ	$\|\boldsymbol{p}\|$	max rcond	w.r.t	min rcond	w.r.t.
1.65×10^{-16}	1.05	2.22	41	0.01272	−1
−0.0494	5.07	8.38	−16	4.44×10^{-15}	41
−0.0494	5.07	8.38	−16	4.44×10^{-15}	41
3.72	1.01	0.0494	29	9.86×10^{-17}	41

Here L is almost singular and yet all eigenvalues are robust. Thus cond(L) is a poor predictor of sensitivity.

(3a) The tridiagonal is W_{21}^{+}. See [13], p. 330.

$$n = 21, \qquad \|L\| = 3.278, \qquad \mathrm{cond}(L) = 10.49,$$

$$\Omega = \mathrm{diag}(++++++++++-++++++++++).$$

λ	$\|p\|$	max rcond	w.r.t	min rcond	w.r.t.
0.254	1.39	1.528	12	3.62×10^{-14}	11
0.948	1.35	1.470	−10	6.34×10^{-12}	−1
−1.13	1.72	3.26	11	1.034×10^{-15}	1
3.04	1.01	0.352	7	2.24×10^{-8}	−1

This well-known matrix has one negative eigenvalue. It is the most sensitive but nevertheless is extremely well determined as is guaranteed by the small value of cond(L).

(3b) W_{21}^{+} shifted by eig(18) = 9.210678647304919.

$$n = 21, \qquad \|L\| = 1810, \qquad \mathrm{cond}(L) = 2.376 \times 10^{11},$$

$$\Omega = \mathrm{diag}(+ - - - - - - - - - + - - - - - - - - - + -).$$

λ	$\|p\|$	max rcond	w.r.t	min rcond	w.r.t.
-2.8×10^{-16}	2.61	5.28	11	0.292	1
5.64×10^{-11}	605	3.66×10^{5}	11	5.18×10^{-6}	21
5.64×10^{-11}	605	3.66×10^{5}	11	5.18×10^{-6}	21
−7.42	1	0.354	14	3.4×10^{-26}	21

(3c) W_{21}^{+} shifted by eig(19) = 9.210678647361332.

$$n = 21, \qquad \|L\| = 3.215, \qquad \mathrm{cond}(L) = 2.405 \times 10^{8}$$

$$\Omega = \mathrm{diag}(+ - + -).$$

λ	$\|p\|$	max rcond	w.r.t	min rcond	w.r.t.
6.44×10^{-16}	0.527	1.708	20	0.292	−1
-5.64×10^{-11}	1	1.0	7	6.34×10^{-6}	21
6.44×10^{-16}	0.527	1.708	20	0.292	−1
−6.17	1	0.348	13	6.38×10^{-24}	21

(3b) and (3c) show the extreme sensitivity to a shift. The shifts agree to 11 decimals, both matrices are almost singular and yet (3c) yields very small condition numbers while (3b) yields very large ones. We have no explanation of this phenomenon.

7. Nonsymmetric tridiagonals

Any real tridiagonal is diagonally similar to ΔT, where T is real symmetric and $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_n)$, $\delta_i = \pm 1$. This similarity transformation is an instance of balancing a nonsymmetric matrix; a simple and often used way to reduce the norm of the original matrix. We assume that the eigenvalues λ_i are distinct but we allow them to be complex. Thus

$$T\boldsymbol{s}_i = \Delta \boldsymbol{s}_i \lambda_i, \quad i = 1, \ldots, n, \tag{17}$$

$$S = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n], \quad \text{possibly complex.} \tag{18}$$

Any such T may be written as

$$T = L\Omega L^{\mathrm{t}}, \tag{19}$$

where L is lower triangular and Ω is a sign symmetric permutation (s.s.p.) matrix. For simplicity we shall assume that Ω is diagonal. A convenient normalization for the eigenvectors is

$$S^{\mathrm{t}}\Delta S = I. \tag{20}$$

Let $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and denote by $\Lambda^{1/2}$ the principal square root of Λ; $\lambda = \rho \exp(2\mathrm{i}\theta)$, $\lambda^{1/2} = \rho \exp(\mathrm{i}\theta)$, $-\pi/2 < \theta \leqslant \pi/2$. As in the symmetric case there is an improper singular value decomposition of L;

$$L = \Delta S \Lambda^{1/2} P^{\mathrm{t}}, \tag{21}$$

defining P which is complex in general. Note that

$$\begin{aligned} P^{\mathrm{t}}\Omega P &= \Lambda^{-1/2}(\Delta S)^{-1} L\Omega L^{\mathrm{t}} (\Delta S)^{-\mathrm{t}} \Lambda^{-1/2}, \quad \text{by Eq.(21)}, \\ &= \Lambda^{-1/2}(\Delta S)^{-1} T (\Delta S)^{-t} \Lambda^{-1/2}, \quad \text{by Eq.(19)}, \\ &= \Lambda^{-1/2}\left[\Lambda S^{-1}(\Delta S)^{-t}\right]\Lambda^{-1/2}, \quad \text{by Eq.(17)} = I, \text{by Eq.(20)}. \end{aligned} \tag{22}$$

With Δ and Ω fixed we study how Λ depends on L. Write

$$L = \mathrm{bidiag}\begin{pmatrix} a_1 & a_2 & \cdot & a_{n-1} & a_n \\ & b_1 & b_2 & \cdot & b_{n-1} & \end{pmatrix}$$

and consider a typical 'singular' triple $(\lambda, \boldsymbol{s}, \boldsymbol{p}^{\mathrm{t}})$;

$$L(\Omega \boldsymbol{p}) = (\Delta s)\lambda^{1/2}, \tag{23}$$

$$\boldsymbol{s}^{\mathrm{t}} L = \lambda^{1/2} \boldsymbol{p}^{\mathrm{t}}. \tag{24}$$

Now we can state the relative sensitivity of the eigenvalues of $\Delta L\Omega L^{\mathrm{t}}$ to L's entries. The expressions may be complex.

Theorem 3. *If ΔT has distinct eigenvalues, and Eq. (19) holds with $\Omega = \mathrm{diag}(\omega_1, \ldots, \omega_n)$, $\omega_i = \pm 1$, then, for $\lambda \neq 0$,*

$$\frac{1}{2}\frac{a_j}{\lambda}\frac{\partial\lambda}{\partial a_j} = \sum_{k=1}^{j}\delta_k s(k)^2 - \sum_{m=1}^{j-1}\omega_m p(m)^2 = \sum_{m=j}^{n}\omega_m p(m)^2 - \sum_{k=j+1}^{n}\delta_k s(k)^2,$$

$$\frac{1}{2}\frac{b_j}{\lambda}\frac{\partial\lambda}{\partial b_j} = \sum_{m=1}^{j}\left[\omega_m p(m)^2 - \delta_m s(m)^2\right] = \sum_{m=j+1}^{n}\left[\delta_m s(m)^2 - \omega_m p(m)^2\right].$$

Proof. Premultiply Eq. (23) by $\boldsymbol{s}^{\mathrm{t}}$ and use Eq. (20) to find

$$\lambda^{1/2} = \boldsymbol{s}^{\mathrm{t}} L\Omega \boldsymbol{p}.$$

Thus $\lambda^{1/2}$ depends linearly on a_j and b_j, the entries of L, and

$$\frac{\partial}{\partial a_j}\left(\lambda^{1/2}\right) = s(j)\omega_j p(j), \tag{25}$$

$$\frac{\partial}{\partial b_j}\left(\lambda^{1/2}\right) = s(j+1)\omega_j p(j). \tag{26}$$

Next write out Eqs. (23) and (24) in detail,

$$b_{j-1}\omega_{j-1}p(j-1) + a_j\omega_j p(j) = \delta_j s(j)\lambda^{1/2}, \quad s(j)a_j + s(j+1)b_j.$$

Multiply the first equation by $s(j)$ and the second one by $\omega_j p(j)$ and substitute the expressions from Eqs. (25) and (26) to find for $j = 1, 2, \ldots, n$ and $b_0 = b_n = 0$,

$$b_{j-1}\frac{\partial}{\partial b_{j-1}}\left(\lambda^{1/2}\right) + a_j\frac{\partial}{\partial a_j}\left(\lambda^{1/2}\right) = \delta_j s(j)\lambda^{1/2}, \tag{27}$$

$$a_j\frac{\partial}{\partial a_j}\left(\lambda^{1/2}\right) + b_j\frac{\partial}{\partial b_j}\left(\lambda^{1/2}\right) = \lambda^{1/2}\omega_j p(j)^2. \tag{28}$$

Finally observe that

$$\left(\frac{a_j}{\lambda^{1/2}}\right)\frac{\partial}{\partial a_j}\left(\lambda^{1/2}\right) = \frac{1}{2}\frac{a_j}{\lambda}\frac{\partial\lambda}{\partial a_j}.$$

The system of equations (27) and (28) is triangular and may be solved by setting $j = 1, 2, \ldots,$ in Eqs. (27) and (28) to give the first expressions for the sensitivities. Solving in reverse order gives the second expressions. □

Theorem 3 shows that an eigenvalue (possible complex) of the pair $L\Omega L^{\mathrm{t}}, \Delta$ is defined to high relative accuracy whenever its singular vectors $\boldsymbol{p}$ and $\boldsymbol{s}$ have modest norms, say $\|\boldsymbol{p}\| < 10$, $\|\boldsymbol{s}\| < 10$. Note that Eq. (20) does not guarantee that $\|\boldsymbol{s}\| = 1$.

Acknowledgements

I would like to thank Inderjit Dhillon for producing the numerical examples in Section 6.

References

[1] B.N. Parlett, The Symmetric Eigenvalue Problem, 2nd edition, SIAM, Philadelphia, 1997.
[2] P. Deift, J.W. Demmel, L.-C. Li, C. Tomei, The bidiagonal singular value decomposition and Hamiltonian mechanics, SIAM J. Numer. Anal. 28 (1991) 1463–1516.
[3] J.W. Demmel, K. Veselić, Jacobi's method is more accurate than QR, SIAM J. Matrix Anal. Appl. 13 (1992) 1204–1245.
[4] I. Slapnicar, Accurate Symmetric Eigenreduction by a Jacobi Method, Thesis, Fernuniversitat Hagen, Germany, 1992.
[5] K. Veselić, I. Slapnicar, Floating point perturbations of Hermitian matrices, Linear Algebra Appl. 195 (1993) 81–116.
[6] J.W. Demmel, W. Kahan, Accurate singular values of bidiagonal matrices, SIAM J. Sci. Stat. Comput. 11 (1990) 873–912.
[7] J. Barlow, J.W. Demmel, Computing accurate eigensystems of scaled diagonally dominant matrices, SIAM J. Numer. Anal. 27 (1990) 762–791.
[8] S.C. Eisenstat, I.C.F. Ipsen, Relative perturbation techniques for singular value problems, SIAM J. Numer. Anal. 32 (1995) 1972–1988.
[9] V. Hari, Z. Drmač, On scaled almost diagonal Hermitian matrix pairs, SIAM J. Matrix. Anal. Appl., to appear.
[10] R.-C. Li, Relative perturbation theory: (I) eigenvalue and singular value variations, Tech. Rep. Univ. of California, Berkeley, UCB/CSD-94-855, Computer Science Division, Department of EECS, Univ. of California, Berkeley, 1994. (Revised January 1996), to appear SIAM Matrix Anal. Appl.
[11] J.W. Demmel, Applied Linear Algebra, SIAM, Philadelphia, 1997.
[12] K. Veselić, On linear vibrational systems with one-dimensional damping, Appl. Anal. 29 (1988) 1–18.
[13] J.H. Wilkinson, The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.